

Strong Consistency of Reduced K -means Clustering

Yoshikazu Terada

*Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama,
Toyonaka, Osaka, Japan*

e-mail: terada@sigmath.es.osaka-u.ac.jp

Abstract: Reduced k -means clustering is a method for clustering objects in a low-dimensional subspace. The advantage of this method is that both clustering of objects and low-dimensional subspace reflecting the cluster structure are simultaneously obtained. In this paper, the relationship between conventional k -means clustering and reduced k -means clustering is discussed. Conditions ensuring almost sure convergence of the estimator of reduced k -means clustering as unboundedly increasing sample size have been presented. The results for a more general model considering conventional k -means clustering and reduced k -means clustering are provided in this paper. Moreover, a new criterion and its consistent estimator are proposed to determine the optimal dimension number of a subspace, given the number of clusters.

Keywords and phrases: clustering, dimension reduction, k -means.

1. Introduction

The aim of cluster analysis is the discovery of a finite number of homogeneous classes from data. In some cases, a cluster structure is considered to lie in a low-dimensional subspace of data, and the following procedure is applied:

- Step 1.** Principal component analysis (PCA) is performed, and the first few components are obtained.
- Step 2.** Conventional k -means clustering is performed for the principal scores on the first few principal components.

This two-step procedure is called “tandem clustering” by Arabie & Hubert (1994) and has been discouraged by several authors (e.g., Arabie & Hubert, 1994; Chang, 1983; De Soete & Carroll, 1994). Because the first few principal components of PCA do not necessarily reflect the cluster structure in data, the appropriate clustering result may not be obtained by using the tandem

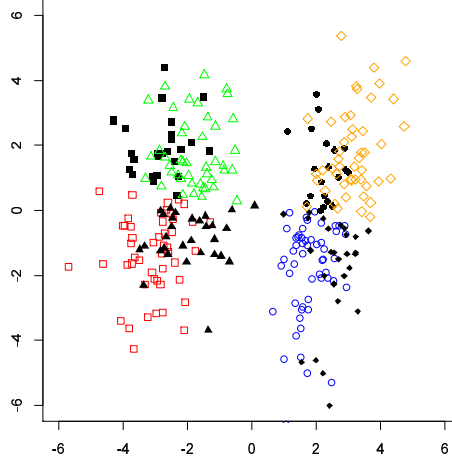


FIG 1. *First two dimensions of the principal component analysis and result of the tandem clustering (Black points represent the misclassification objects).*

clustering approach. Figure 1 shows that the first two principal components do not reflect the cluster structure, and the clustering result of the tandem clustering is incorrect. De Soete & Carroll (1994) proposed reduced k -means (RKM) clustering. RKM clustering simultaneously determines the clusters of objects on the basis of the k -means criterion and the subspace that is informative about the cluster structure in data on the basis of component analysis. In other words, for given data points $\mathbf{x}_1, \dots, \mathbf{x}_n$ in \mathbb{R}^p , the fixed cluster number k and the dimension number of subspace q ($q < \min\{k - 1, p\}$), RKM clustering is defined by the minimization problem of the following loss function:

$$RKM_n := \frac{1}{n} \sum_{i=1}^n \min_{1 \leq j \leq k} \|\mathbf{x}_i - A\mathbf{f}_j\|^2, \quad (1)$$

where $\mathbf{f}_j \in \mathbb{R}^q$ and A is a $p \times q$ columnwise orthonormal matrix. For some clustering methods related to k -means clustering, several authors have discussed their statistical properties (e.g., Abraham et al., 2003; García-Escudero et al., 1999; Pollard, 1981; Pollard, 1982; von Luxburg et al., 2008). However, because RKM clustering is proposed in the framework of descriptive statistics,

the statistical properties are not discussed. When data points are independently drawn from a population distribution P , the objective function is rewritten as

$$RKM(F, A, P_n) := \int \min_{\mathbf{f} \in F} \|\mathbf{x} - A\mathbf{f}\| P_n(d\mathbf{x}),$$

where F is a set containing k or fewer points in \mathbb{R}^q , and P_n is the empirical measure obtained from the data. For each fixed F and A , the strong law of large numbers (SLLN) shows that

$$\lim_{n \rightarrow \infty} RKM(F, A, P_n) = RKM(F, A, P) := \int \min_{\mathbf{f} \in F} \|\mathbf{x} - A\mathbf{f}\| P(d\mathbf{x}) \quad \text{a.s.}$$

Thus, we wish to ensure that the global minimizer of $RKM(\cdot, \cdot, P_n)$ converges almost surely to the global minimizers of $RKM(\cdot, \cdot, P)$, say the population global minimizers.

In this paper, the strong consistency of RKM under i.i.d. sampling is proven. For this purpose, the framework of the proof of the strong consistency of the k -means clustering approach proposed by Pollard (1981) is used; in this framework, the existence and uniqueness of the population global minimizers are assumed for consistency. Conditions for the existence of the global minimizers are not discussed. For RKM clustering, the uniqueness of the population global minimizers cannot be assumed because RKM clustering has rotational indeterminacy. Therefore, the sufficient condition for the existence of the population global minimizers must be derived; it is also necessary to establish that the distance between the sample estimator and the *set* of global minimizers converges almost surely to zero, as the sample size approaches infinity.

This paper is organized as follows. In Section 2, the original algorithm of RKM clustering and visualization of the result are described. Then, the relationship between the conventional k -means clustering method and RKM clustering is presented. The notation and some properties of RKM, including the rotational indeterminacy, is introduced in Section 3. The uniform SLLN and continuity of the objective function of RKM clustering are presented in Section 4. In Section 5, conditions for the existence of the population global minimizers are determined, and a theorem regarding the strong consistency of RKM clustering is stated. In Section 6, the main proof of the consistency theorem is explained. In Section 7, a new criterion and its consistent estimator are proposed to determine the optimal dimension number of a subspace, given

the number of clusters. Moreover, the effectiveness of the criterion through numerical experiments are illustrated.

2. Reduced k -means clustering

2.1. Algorithm and visualization of reduced k -means clustering

Let $X = (x_{ij})_{n \times p}$ be a data matrix and \mathbf{x}_i ($i = 1, \dots, n$) be row vectors of X , where n is the number of objects and p is the number of variables. The number of clusters and components to which the variables are reduced are denoted by k and q , respectively. RKM clustering is defined as the minimizing problem of the following criterion:

$$RKM_n(A, F, U \mid k, q) := \|X - UFA^T\|_F^2 = \sum_{i=1}^n \min_{1 \leq j \leq k} \|\mathbf{x}_i - A\mathbf{f}_j\|^2, \quad (2)$$

where $\|\cdot\|$ and $\|\cdot\|_F$ denote the usual Euclidean norm and Frobenius norm, respectively, $U = (u_{ij})_{n \times k}$ is a binary membership matrix that specifies cluster membership for each objects, $A = (a_{ij})_{p \times q}$ is a column-wise orthonormal loading matrix, $F = (f_{ij})_{k \times q}$ is a centroid matrix, and \mathbf{f}_j is a centroid of the j th cluster for each $j = 1, \dots, k$. For example, this problem can be solved by the following alternating least square algorithm:

Step 0. First, initial values are chosen for A , F , and U .

Step 1. $Q\Sigma P^T$ is expressed as the singular value decomposition of $(UF)^T X$, where Q is a $q \times q$ orthonormal matrix, Σ is a $q \times q$ diagonal matrix, and P is a $p \times q$ columnwise orthonormal matrix. A is updated by PQ^T .

Step 2. For each $i = 1, \dots, n$ and each $j = 1, \dots, k$, we update u_{ij} by

$$u_{ij} = \begin{cases} 1 & \text{iff } \|A^T \mathbf{x}_i - \mathbf{f}_j\|^2 < \|A^T \mathbf{x}_i - \mathbf{f}_{j'}\|^2 \text{ for each } j' \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Step 3. F is updated using $(U^T U)^{-1} U^T X A$.

Step 4. Finally, the value of the function RKM_n for the present values of A , F , and U is computed. When the present values have decreased the function value, A , F , and U are update in accordance with Steps 1–3. Otherwise, the algorithm has converged.

Other formulations and algorithms for RKM clustering have been presented by De Soete & Carrol (1994) and Timmerman et al. (2010).

The algorithms for RKM clustering monotonically decrease the function RKM_n . As shown below, because RKM_n is bounded, the solution for each iteration converges to a local minimum point. Because of the binary constraint on U , the solutions of these algorithms may often be local minimums. To prevent this, many random starts are required to be used.

The objective function RKM_n can be decomposed into two terms:

$$RKM_n(A, F, U \mid k, q) = \|X - XAA^T\|_F^2 + \|XA - UF\|_F^2. \quad (3)$$

The first term of equation (3) is the objective function of the PCA, and the second term is the k -means criterion in a low dimensional subspace. Thus, for optimal solutions \hat{A} , \hat{F} , and \hat{U} , we have $\hat{F} = (\hat{U}^T \hat{U})^{-1} \hat{U}^T \hat{X} \hat{A}$. Using the optimal solutions \hat{A} , \hat{F} , and \hat{U} , the low-dimensional representation of the objects and cluster centers can be obtained:

$$Y := X\hat{A} \quad \text{and} \quad G := (\hat{U}^T \hat{U})^{-1} \hat{U}^T Y. \quad (4)$$

Using Y and \hat{A} , a biplot reflecting the cluster structure can be presented. Figure 2 shows the biplot of the RKM clustering for the same data as that used in Figure 1.

2.2. The relationship between the conventional k -means and the RKM clusterings

The objective function of the conventional k -means clustering method is given by

$$KM_n(C, U \mid k) := \|X - UC\|_F^2, \quad (5)$$

where C is an $k \times p$ cluster center matrix. $P\Sigma Q^T$ is expressed as the singular value decomposition of C , where P is an $k \times k$ orthonormal matrix, Σ is an $k \times k$ diagonal matrix, and Q is a $p \times k$ column-wise orthonormal matrix. Function (5) can be expressed as

$$\|X - UC\|^2 = \|X - UP\Sigma Q^T\|_F^2.$$

Considering $P\Sigma$ and Q as a low-dimensional centroid matrix F and a loading matrix A , respectively, function (5) is equivalent to the objective function of RKM, $RKM_n(A, F, U \mid k, k)$. Thus, RKM clustering includes the conventional k -means clustering analysis as a special case.

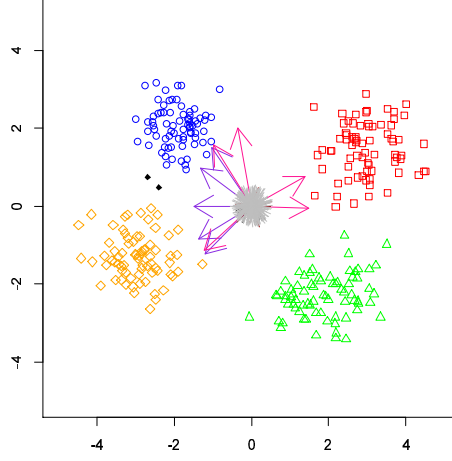


FIG 2. Biplot of the result of RKM clustering for the same data set as Figure 1 (Black points represent the misclassification objects).

3. Preliminaries

Let (Ω, \mathcal{F}, P) be a probability space and $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent random variables with a common population distribution P on \mathbb{R}^p ; let P_n be the empirical measure based on $\mathbf{X}_1, \dots, \mathbf{X}_n$. For typographical convenience, the set of all $p \times q$ column-wise orthonormal matrices are denoted by $\mathcal{O}(p \times q)$, and $\mathcal{R}_k := \{R \subset \mathbb{R}^q \mid \#(R) \leq k\}$, where $\#(R)$ is the cardinality of R . Thus, the parameter space is denoted by $\Xi_k := \mathcal{R}_k \times \mathcal{O}(p \times q)$. $B_q(r)$ denotes the q -dimensional closed ball of radius r centered at the origin. For each $M > 0$, define $\mathcal{R}_k^*(M) := \{R \subset B_q(M) \mid \#(R) \leq k\}$ and $\Theta_k^*(M) := \mathcal{R}_k^*(M) \times \mathcal{O}(p \times q)$. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be a non-negative decreasing function and Q be a probability measure on \mathbb{R}^p . For each finite subset $F \subset \mathbb{R}^q$ and each $A \in \mathcal{O}(p \times q)$, the loss function of RKM with Q is defined by

$$\Phi(F, A, Q) := \int \min_{\mathbf{f} \in F} \phi(\|\mathbf{x} - A\mathbf{f}\|) Q(d\mathbf{x}).$$

Write

$$m_k(Q) := \inf_{(F, A) \in \Xi_k} \Phi(F, A, Q) \quad \text{and} \quad m_k^*(Q \mid M) := \inf_{(F, A) \in \Theta_k^*(M)} \Phi(F, A, Q).$$

For $\theta = (F, A) \in \Xi_k$, both descriptions $\Phi(\theta, Q)$ and $\Phi(F, A, Q)$ are used. In addition, $\Theta' := \{\theta \in \Xi_k \mid m_k(P) = \Phi(\theta, P)\}$ and $\Theta'_n := \{\theta \in \Xi_k \mid m_k(P_n) = \Phi(\theta, P_n)\}$. For each $M > 0$, $\Theta^* := \{\theta \in \Theta'_k(M) \mid m_k^*(P_n \mid M) = \Phi(\theta, P_n)\}$ and $\Theta_n^* := \{\theta \in \Theta'_k(M) \mid m_k^*(P_n \mid M) = \Phi(\theta, P_n)\}$. The parameters $\Theta'(k)$ and $\Theta'_n(k)$ are used to emphasize that Θ' and Θ'_n are dependent on the index k . One of the measurable estimators in Θ'_n will be denoted by $\hat{\theta}_n$ or $\hat{\theta}_n(k)$. Similarly, we will also denote one of the measurable estimators in Θ_n^* by $\hat{\theta}_n^*$ or $\hat{\theta}_n^*(k)$. To illustrate the existence of measurable estimators, see Section 6.7 of Pfanzagl (1996).

Let $d_F(\cdot, \cdot)$ be the distance between two matrices based on Frobenius norm and $d_H(\cdot, \cdot)$ the Hausdorff distance, which is defined for finite subsets $A, B \subset \mathbb{R}^q$ as

$$d_H(A, B) := \max_{\mathbf{a} \in A} \left\{ \min_{\mathbf{b} \in B} \|\mathbf{a} - \mathbf{b}\| \right\}.$$

Moreover, let d be the product distance with d_F and d_H . In this paper, the distance between $\hat{\theta}_n$ and Θ' is defined as

$$d(\hat{\theta}_n, \Theta') := \inf\{d(\hat{\theta}_n, \theta) \mid \theta \in \Theta'\}.$$

To clarify the minimization procedures, the function ϕ must satisfy some regularity conditions. As proposed by Pollard (1981), it is assumed that ϕ is continuous, and $\phi(0) = 0$. Moreover, to control the growth of ϕ , it is assumed that

$$\exists \lambda > 0; \forall r > 0; \phi(2r) \leq \lambda \phi(r).$$

For each $\mathbf{f} \in \mathbb{R}^q$ and each $A \in \mathcal{O}(p \times q)$,

$$\begin{aligned} \int \phi(\|\mathbf{x} - A\mathbf{f}\|)P(d\mathbf{x}) &\leq \int \phi(\|\mathbf{x}\| + \|A\mathbf{f}\|)P(d\mathbf{x}) = \int \phi(\|\mathbf{x}\| + \|\mathbf{f}\|)P(d\mathbf{x}) \\ &= \int_{\|\mathbf{f}\| > \|\mathbf{x}\|} \phi(2\|\mathbf{f}\|)P(d\mathbf{x}) + \int_{\|\mathbf{f}\| \leq \|\mathbf{x}\|} \phi(2\|\mathbf{x}\|)P(d\mathbf{x}) \\ &\leq \phi(2\|\mathbf{f}\|) + \lambda \int \phi(\|\mathbf{x}\|)P(d\mathbf{x}). \end{aligned}$$

Therefore, as long as $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x})$ is finite, $\Phi(F, A, P)$ is also finite for each F and each $A \in \mathcal{O}(p \times q)$.

Let R be a $q \times q$ orthonormal matrix, i.e., $R^T R = R R^T = I_q$. For each $\mathbf{f} \in \mathbb{R}^q$ and each $A \in \mathcal{O}(p \times q)$,

$$\int \phi(\|\mathbf{x} - A\mathbf{f}\|)P(d\mathbf{x}) = \int \phi(\|\mathbf{x} - AR^T R\mathbf{f}\|)P(d\mathbf{x}).$$

It follows that Θ' is not a singleton when $\Theta' \neq \emptyset$, thus suggesting that RKM clustering has rotational indeterminacy.

4. The uniform SLLN and the continuity of $\Phi(\cdot, \cdot, P)$

Proposition 1. *Let $M > 0$ be an arbitrary number. Let \mathcal{G} denote the class of all P -integrable functions on \mathbb{R}^p of the form*

$$g_{(F, A)}(\mathbf{x}) := \min_{\mathbf{f} \in F} \phi(\|\mathbf{x} - A\mathbf{f}\|),$$

where (F, A) takes all values over $\Theta_k^*(M)$. Suppose that $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \infty$. Then,

$$\lim_{n \rightarrow \infty} \sup_{g \in \mathcal{G}} \left| \int g(\mathbf{x}) P_n(d\mathbf{x}) - \int g(\mathbf{x}) P(d\mathbf{x}) \right| = 0 \quad \text{a.s.} \quad (6)$$

Proof. DeHardt (1971) provided the sufficient condition for the uniform SLLN (6); for all $\epsilon > 0$, there exists a finite class of functions \mathcal{G}_ϵ such that for each $g \in \mathcal{G}$, \dot{g} and \bar{g} exist in \mathcal{G}_ϵ with $\dot{g} \leq g \leq \bar{g}$ and $\int \bar{g}(\mathbf{x}) P(d\mathbf{x}) - \int \dot{g}(\mathbf{x}) P(d\mathbf{x}) < \epsilon$.

An arbitrary $\epsilon > 0$ is selected, and $S_{p \times q}(\sqrt{q})$ denotes the surface of the sphere on $\mathbb{R}^{p \times q}$ of radius \sqrt{q} centered at the origin. To find such a finite class \mathcal{G}_ϵ , D_{δ_1} is defined as the finite set of \mathbb{R}^q satisfying

$$\forall \mathbf{f} \in B_q(M); \exists \mathbf{g} \in D_{\delta_1}; \|\mathbf{f} - \mathbf{g}\| < \delta_1$$

and $\mathcal{A}_{p \times q, \delta_2}$ as the finite sets of $S_{p \times q}(\sqrt{q})$ satisfying

$$\forall A \in S_{p \times q}(\sqrt{q}); \exists B \in \mathcal{A}_{p \times q, \delta_2}; \|A - B\|_F < \delta_2.$$

Define $\mathcal{R}_{k, \delta_1} := \{F \in \mathcal{R}_k^*(M) \mid F \subset D_{\delta_1}\}$. Take \mathcal{G}_ϵ as the finite class of functions of the form

$$\min_{\mathbf{f} \in F'} \phi(\|\mathbf{x} - A'\mathbf{f}\| + \sqrt{q}\delta_1 + M\delta_2) \quad \text{or} \quad \min_{\mathbf{f} \in F'} \phi(\|\mathbf{x} - A'\mathbf{f}\| - \sqrt{q}\delta_1 - M\delta_2),$$

where (F', A') takes all values over $\mathcal{R}_{k, \delta_1} \times \mathcal{A}_{p \times q, \delta_2}$ and $\phi(r)$ is defined as zero for all negative $r < 0$.

For given $F = \{\mathbf{f}_1, \dots, \mathbf{f}_k\} \in \mathcal{R}_k^*(M)$ and $A \in \mathcal{O}(p \times q)$, there exists $F' = \{\mathbf{f}'_1, \dots, \mathbf{f}'_k\} \in \mathcal{R}_{k, \delta_1}$ with $\|\mathbf{f}_i - \mathbf{f}'_i\| < \delta_1$ for each i and each $A' \in \mathcal{A}_{p \times q, \delta_2}$ with $\|A - A'\|_F < \delta_2$. Corresponding to each $g_{(F, A)} \in \mathcal{G}$, choose

$$\bar{g}_{(F, A)} := \min_{\mathbf{f} \in F'} \phi(\|\mathbf{x} - A'\mathbf{f}\| + \sqrt{q}\delta_1 + M\delta_2)$$

and

$$\dot{g}_{(F, A)} := \min_{\mathbf{f} \in F'} \phi(\|\mathbf{x} - A'\mathbf{f}\| - \sqrt{q}\delta_1 - M\delta_2).$$

Because ϕ is a monotone function and

$$\|\mathbf{x} - A'\mathbf{f}'_i\| - \sqrt{q}\delta_1 - M\delta_2 \leq \|\mathbf{x} - A\mathbf{f}_i\| \leq \|\mathbf{x} - A'\mathbf{f}'_i\| + \sqrt{q}\delta_1 + M\delta_2$$

for each i and each $\mathbf{x} \in \mathbb{R}^p$, these functions ensure that $\dot{g}_{(F, A)} \leq g_{(F, A)} \leq \bar{g}_{(F, A)}$.

If we choose $R > 0$ to be greater than $\sqrt{q}\delta_1 + M\delta_2 + M\sqrt{q}$,

$$\begin{aligned} & \int [\bar{g}_{(F, A)}(\mathbf{x}) - \dot{g}_{(F, A)}(\mathbf{x})] P(d\mathbf{x}) \\ & \leq \int \sum_{i=1}^k \left[\phi(\|\mathbf{x} - A'\mathbf{f}'_i\| + \sqrt{q}\delta_1 + M\delta_2) \right. \\ & \quad \left. - \phi(\|\mathbf{x} - A'\mathbf{f}'_i\| - \sqrt{q}\delta_1 - M\delta_2) \right] P(d\mathbf{x}) \\ & \leq k \sup_{\|\mathbf{x}\| \leq R} \sup_{\mathbf{f} \in B(5M)} \sup_{A \in S_{p \times q}(\sqrt{q})} \left[\phi(\|\mathbf{x} - A\mathbf{f}\| + \sqrt{q}\delta_1 + M\delta_2) \right. \\ & \quad \left. - \phi(\|\mathbf{x} - A\mathbf{f}\| - \sqrt{q}\delta_1 - M\delta_2) \right] + 2k\lambda \int_{\|\mathbf{x}\| \geq R} \phi(\|\mathbf{x}\|) P(d\mathbf{x}). \end{aligned}$$

The second term would be less than $\epsilon/2$ if R is sufficiently large. Moreover, because ϕ is uniform continuous on a bounded set, the first term can be less than $\epsilon/2$ if $\delta_1, \delta_2 > 0$ is sufficiently small. Thus, the uniform SLLN is proven. \square

Similarly, the continuity of $\Phi(\cdot, P)$ on $\Theta_k^*(M)$ can be proven.

Proposition 2. *Let $M > 0$ be an arbitrary number. Suppose that $\int \phi(\|\mathbf{x}\|) P(d\mathbf{x})$. Then, $\Phi(\cdot, P)$ is continuous on $\Theta_k^*(M)$.*

Proof. If $(F, A), (G, B) \in \Theta_k^*$ are select such that $d_H(F, G) < \delta_1$ and $\|A - B\|_F < \delta_2$, then for each $\mathbf{g} \in G$, there exists $\mathbf{g}(\mathbf{f}) \in F$ with $\|\mathbf{g} - \mathbf{g}(\mathbf{f})\| < \delta_1$, and furthermore,

$$\begin{aligned} & \Phi(F, \mathbf{A}, P) - \Phi(G, \mathbf{B}, P) \\ & = \int \left[\min_{\mathbf{f} \in F} \phi(\|\mathbf{x} - \mathbf{A}\mathbf{f}\|) - \min_{\mathbf{g} \in G} \phi(\|\mathbf{x} - \mathbf{B}\mathbf{g}\|) \right] P(d\mathbf{x}) \end{aligned}$$

$$\begin{aligned}
&\leq \int \max_{\mathbf{g} \in G} [\phi(\|\mathbf{x} - \mathbf{A}\mathbf{f}(\mathbf{g})\|) - \phi(\|\mathbf{x} - \mathbf{B}\mathbf{g}\|)] P(d\mathbf{x}) \\
&\leq \int \sum_{\mathbf{g} \in G} [\phi(\|\mathbf{x} - \mathbf{B}\mathbf{g}\| + M\delta_2 + \delta_1) - \phi(\|\mathbf{x} - \mathbf{B}\mathbf{g}\|)] P(d\mathbf{x}) \\
&\leq k \sup_{\|\mathbf{x}\| \leq R} \max_{\mathbf{g} \in G} [\phi(\|\mathbf{x} - \mathbf{B}\mathbf{g}\| + M\delta_2 + \delta_1) - \phi(\|\mathbf{x} - \mathbf{B}\mathbf{g}\|)] \\
&\quad + 2k\lambda \int_{\|\mathbf{x}\| \geq R} \phi(\|\mathbf{x}\|) P(d\mathbf{x}) \tag{7}
\end{aligned}$$

for $R > \delta_1 + M(1 + \delta_2)$. When a sufficiently large R and a sufficiently small $\delta_1, \delta_2 > 0$ are selected, the last bound is less than ϵ . For each $\mathbf{f} \in F$, there also exists $\mathbf{f}(\mathbf{g}) \in G$ with $\|\mathbf{f} - \mathbf{f}(\mathbf{g})\| < \delta_1$. Therefore, the other inequality necessary for the continuity is obtained by interchanging (F, A) and (G, B) in the inequality (7). \square

5. The consistency theorem

5.1. The existence of the population global optimizers

The aim of this paper is to prove that, for a fixed measure P satisfying some natural assumptions, the infimum distance between the (measurable) estimator $\hat{\theta}_n$ with $\Phi(\hat{\theta}_n) = m_k(P_n)$ and parameters achieving $m_k(P)$ converges almost surely to 0, as the sample size goes to infinity. However, there may be no such parameters. Thus, before providing the consistency theorem, the sufficient condition for the existence of parameters achieving $m_k(P)$ in Ξ_k is provided. The following proposition ensures the existence of such parameters. The proof and some details about the proposition are given in Appendix A.

Proposition 3. *Suppose that $\int \phi(\|\mathbf{x}\|) P(d\mathbf{x}) < \infty$ and that $m_j(P) > m_k(P)$ for $j = 1, 2, \dots, k-1$. Then, $\Theta' \neq \emptyset$.*

From Lemma 4 in Appendix A, there exists $M > 0$ such that $F \subset B_q(5M)$ for all $(F, A) \in \Theta'$. Moreover, under the assumption of Proposition 3, the following identification condition can be proven:

$$\inf_{\theta \in \Theta_k^*(5M): d(\theta, \Theta') \geq \epsilon} \Phi(\theta, P) > \inf_{\theta \in \Theta'} \Phi(\theta, P) \quad \text{for all } \epsilon > 0.$$

The proof of the identification condition is also given in Appendix A. The identification condition is used in Section 6.

5.2. Strong consistency of reduced k -means clusterings

If the parameter space is $\Theta_k^*(M)$, the strong consistency of RKM clustering can be proven. Note that since $\Theta_k^*(M)$ is compact, we have $\Theta^* \neq \emptyset$ and the identification condition:

$$\inf_{\theta \in \Theta_\epsilon^*(M)} \Phi(\theta, P) > \inf_{\theta \in \Theta^*} \Phi(\theta, P) \quad \text{for all } \epsilon > 0,$$

where $\Theta_\epsilon^*(M) := \{\theta \in \Theta_k^*(M) \mid d(\theta, \Theta^*) \geq \epsilon\}$.

Proposition 4. *Suppose that $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \infty$. Then, for each $M > 0$,*

$$\lim_{n \rightarrow \infty} d(\hat{\theta}_n^*, \Theta^*) = 0 \quad \text{a.s., and} \quad \lim_{n \rightarrow \infty} m_k^*(P_n \mid M) = m_k^*(P \mid M) \quad \text{a.s.}$$

Proof. Since the uniform SLLN and the continuity of $\Phi(\cdot, P)$, the proof of this proposition is given by the similar argument of the proof of the following consistency theorem. \square

In a study by Pollard (1981), the uniqueness of the parameter is also assumed for the strong consistency theorem. As discussed in Section 3, we cannot assume the uniqueness condition. Thus, the condition that $m_j(P) > m_k(P)$ for $j = 1, 2, \dots, k-1$ is assumed instead of the uniqueness condition.

This condition is equivalent to the distinctness condition that $F(k)$ has k distinct points for all $(F(k), A(k)) \in \Theta'(k)$. Indeed, suppose that there exists $\theta = (F(k), A(k)) \in \Theta'(k)$ such that $F(k)$ have $k-1$ or fewer distinct points; that is, $\#(F(k)) < k$. There exists $i \in \mathbb{N}$ such that $i < k$ and $\theta \in \Xi_k$. Then, $m_i(P) = m_k(P)$, which contradicts to $m_i(P) > m_k(P)$. Thus, the condition that $m_j(P) > m_k(P)$ for $j = 1, 2, \dots, k-1$ implies the distinctness condition. Moreover, this condition is equivalent to $m_{k-1}(P) > m_k(P)$ since $m_k(P) \geq m_l(P)$ for each $k, l \in \mathbb{N}$ satisfying $k < l$.

The following main theorem gives the sufficient condition for the strong consistency of the estimator of RKM clustering.

Theorem 1. *Suppose that $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \infty$ and that $m_j(P) > m_k(P)$ for $j = 1, 2, \dots, k-1$. Then, $\Theta' \neq \emptyset$,*

$$\lim_{n \rightarrow \infty} d(\hat{\theta}_n, \Theta') = 0 \quad \text{a.s., and} \quad \lim_{n \rightarrow \infty} m_k(P_n) = m_k(P) \quad \text{a.s.}$$

6. Proof of Theorem 1

Because almost sure convergence is dealt with, null sets of elements exists for which the convergence does not hold. Hereafter, Ω_1 denotes the set obtained by avoiding a proper null set from Ω . In the first step of the proof, when n is sufficiently large, the estimators of the cluster centers are contained within a compact ball that does not depend on $\omega \in \Omega$. For convenience, it is assumed that $\phi(r) \rightarrow \infty$ as $r \rightarrow \infty$. When ϕ is bounded, the proof is a little complicated.

First, we prove the following lemma.

Lemma 1. *Suppose that $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \infty$. Then, there exists $M > 0$ such that*

$$P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{\omega \mid \forall (F_m, A_m) \in \Theta'_m; F_m(\omega) \cap B_q(M) \neq \emptyset\}\right) = 1.$$

Proof. Select an appropriate value $r > 0$ to satisfy the condition that the ball $B_p(r)$ has positive P measure, i.e., $P(B_p(r)) > 0$. Let M be sufficiently large for satisfying $M > r$ and

$$\phi(M - r)P(B_p(r)) > \int \phi(\|\mathbf{x}\|)P(d\mathbf{x}). \quad (8)$$

From the definition of $\hat{\theta}_n = (F_n, A_n)$, $\Phi(F_n, A_n, P) \leq \Phi(F_0, A, P)$ for any set F_0 containing at most k points and any $A \in \mathcal{O}(p \times q)$. The parameter F_0 is chosen such that it only consists of the origin. Then, by SLLN,

$$\Phi(F_0, A, P_n) = \int \phi(\|\mathbf{x}\|)P_n(d\mathbf{x}) \rightarrow \int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) \quad \text{a.s.},$$

for each $A \in \mathcal{O}(p \times q)$.

Let $\Omega' := \{\omega \in \Omega_1 \mid \forall n \in \mathbb{N}; \exists m \geq n; \exists (F_m, A_m) \in \Theta'_m; F_m(\omega) \cap B_q(M) = \emptyset\}$. By the axiom of choice, for an arbitrary $\omega \in \Omega'$ there exists a subsequence $\{n_l\}_{l \in \mathbb{N}}$ such that $n_s < n_t$ ($s < t$) and $F_{n_l} \cap B_q(M) = \emptyset$. Thus,

$$\begin{aligned} \limsup_l \Phi(F_{n_l}, A_{n_l}, P_{n_l}) &\geq \limsup_l \frac{1}{n_l} \sum_{i \in \{i \mid \mathbf{X}_i \in B_p(r)\}} \min_{1 \leq j \leq k} \phi(\|\mathbf{X}_i - A_{n_l} \mathbf{f}_j\|) \\ &\geq \limsup_l \frac{1}{n_l} \sum_{i \in \{i \mid \mathbf{X}_i \in B_p(r)\}} \phi(M - r) \end{aligned}$$

$$= \phi(M - r) \limsup_l P_{n_l}(B_p(r)) = \phi(M - r)P(B_p(r)).$$

On the other hand, $\limsup_l m_k(P_{n_l}) \leq \lim_l \Phi(F_0, A, P_{n_l})$ because $m_k(P_{n_l}) \leq \Phi(F_0, A, P_{n_l})$. Therefore, we have $\limsup_l m_k(P_{n_l}) \leq \int \phi(\|\mathbf{x}\|)P(d\mathbf{x})$ and $\limsup_l \Phi(F_{n_l}, A_{n_l}, P_{n_l}) > \int \phi(\|\mathbf{x}\|)P(d\mathbf{x})$, which is a contradiction. Therefore, $P(\Omega') = 0$, that is,

$$P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{\omega \mid \forall (F_m, A_m) \in \Theta'_m; F_m(\omega) \cap B_q(M) \neq \emptyset\}\right) = 1.$$

□

Without loss of generality, all F_n can be assumed contain at least one point of $B_q(M)$ when n is sufficiently large. The next lemma shows that for sufficiently large n , there exists $M > 0$ such that the closed ball $B_q(5M)$ contains all estimators of centers. When $k = 1$, the next lemma is obviously satisfied.

From the results in Section 4 and using the same arguments in the final part of this section, the conclusions of the theorem are proven when $k = 1$.

Lemma 2. *Under the assumption of the theorem, there exists $M > 0$ such that*

$$P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{\omega \mid \forall (F_m, A_m) \in \Theta'_m; F_m(\omega) \subset B_q(5M)\}\right) = 1.$$

Proof. Choose $M > 0$ sufficiently large to satisfy the inequality (8) and

$$\lambda \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \epsilon, \quad (9)$$

where $\epsilon > 0$ is selected to ensure $\epsilon + m_k(P) < m_{k-1}(P)$. Note that $m_j(P) \leq m_j^*(P \mid M)$ for $j \in \mathbb{N}$.

Suppose that F_n contains at least one center outside $B_q(5M)$ and consider the effect on $\Phi(F_n, A, P_n)$ by deleting such outside centers from F_n for all $A \in \mathcal{O}(p \times q)$. From Lemma 1, all F_n contain at least one center on $B_q(M)$ when n is sufficiently large, say \mathbf{f}_1 . In the worst case, the cluster of $\mathbf{f}_1 \in B_q(M)$ should contain all sample points belonging to clusters outside $B_q(5M)$. Because these points must be outside $B(2M)$, the increment of

$\Phi(F_n, A, P_n)$ due to the deletion of centers outside $B_q(5M)$ from F_n would be at most

$$\begin{aligned} \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x} - A\mathbf{f}_1\|) P_n(d\mathbf{x}) &\leq \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\| + \|\mathbf{f}_1\|) P_n(d\mathbf{x}) \\ &\leq \int_{\|\mathbf{x}\| \geq 2M} \phi(2\|\mathbf{x}\|) P_n(d\mathbf{x}) \\ &\leq \lambda \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\|) P_n(d\mathbf{x}). \end{aligned}$$

Denote the set obtained by deleting centers outside $B_q(5M)$ from F_n by F_n^* . For each $A \in \mathcal{O}(p \times q)$, (F_n^*, A) is contained in $\Theta_{k-1}^*(5M)$, and thus,

$$\Phi(F_n^*, A, P_n) \geq m_{k-1}^*(P_n \mid 5M) \geq m_{k-1}(P_n).$$

For each \mathbf{x} satisfying $\|\mathbf{x}\| < 2M$ and each $A \in \mathcal{O}(p \times q)$, we have

$$\|\mathbf{x} - A\mathbf{f}\| > 3M \quad \text{for all } \mathbf{f} \notin B_q(5M)$$

and

$$\|\mathbf{x} - A\mathbf{g}\| < 3M \quad \text{for all } \mathbf{g} \in B_q(M).$$

Thus,

$$\int_{\|\mathbf{x}\| < 2M} \min_{\mathbf{f} \in F_n} \phi(\|\mathbf{x} - A\mathbf{f}\|) P_n(d\mathbf{x}) = \int_{\|\mathbf{x}\| < 2M} \min_{\mathbf{f} \in F_n^*} \phi(\|\mathbf{x} - A\mathbf{f}\|) P_n(d\mathbf{x}).$$

for all $A \in \mathcal{O}(p \times q)$. Note that

$$\lim_{n \rightarrow \infty} m_{k-1}^*(P_n \mid 5M) = m_{k-1}^*(P \mid 5M) \quad \text{a.s.}$$

by Proposition 4.

Let $\Omega^* := \{\omega \in \Omega_1 \mid \forall n \in \mathbb{N}; \exists m \geq n; \exists (F_m, A_m) \in \Theta'_m; F_m(\omega) \not\subset B_q(5M)\}$. By the axiom of choice, for an arbitrary $\omega \in \Omega^*$ there exists a subsequence $\{n_l\}_{l \in \mathbb{N}}$ such that $n_s < n_t$ ($s < t$) and $F_{n_l}(\omega) \not\subset B_q(5M)$. For any F with k or fewer points and any $A \in \mathcal{O}(p \times q)$,

$$\begin{aligned} m_{k-1}^*(P \mid 5M) &\leq \liminf_i \Phi(F_{n_i}^*, A_{n_i}, P_{n_i}) \leq \limsup_i \Phi(F_{n_i}^*, A_{n_i}, P_{n_i}) \\ &= \limsup_i \left[\int_{\|\mathbf{x}\| < 2M} \min_{\mathbf{f} \in F_{n_i}} \phi(\|\mathbf{x} - A_{n_i}\mathbf{f}\|) P_{n_i}(d\mathbf{x}) \right] \end{aligned}$$

$$\begin{aligned}
& + \int_{\|\mathbf{x}\| \geq 2M} \min_{\mathbf{f} \in F_{n_i}^*} \phi(\|\mathbf{x} - A_{n_i} \mathbf{f}\|) P_{n_i}(d\mathbf{x}) \Big] \\
& \leq \limsup_n \left[\Phi(F_n, A_n, P_n) + \lambda \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\|) P_n(d\mathbf{x}) \right] \\
& \leq \limsup_n \Phi(F, A, P_n) + \lambda \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\|) P(d\mathbf{x}). \quad (10)
\end{aligned}$$

Set $(F, A) \in \Theta'$; that is, $m_k(P) = \Phi(F, A, P)$. From the requirement of $M > 0$ in the inequality (9) and SLLN, the last bound of the inequality (10) is less than

$$\Phi(F, A, P) + \epsilon = m_k(P) + \epsilon < m_{k-1}(P).$$

This is a contradiction. Thus, the following is obtained

$$P \left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{ \omega \mid \forall (F_m, A_m) \in \Theta'_m; F_m(\omega) \subset B_q(5M) \} \right) = 1.$$

□

For sufficiently large n , all F_n values satisfying

$$\inf_{A \in \mathcal{O}(p \times q)} \Phi(F_n, A, P_n) = m_k(P_n)$$

lie in $\mathcal{R}_k^*(5M)$. From Proposition 3 and Lemma 4, $\mathcal{R}_k^*(5M)$ contains all optimal sets satisfying

$$\inf_{A \in \mathcal{O}(p \times q)} \Phi(F, A, P) = m_k(P).$$

It also follows that Pollard (1981) assume that it is large enough to satisfy that $\mathcal{R}_k^*(5M)$ contains the optimal cluster centers, as the requirement on M , but this requirement is also unnecessary.

In a similar way of Theorem 5.14 (van der Vaart, 1998), if we obtain the continuity of $\Phi(\cdot, \cdot, P)$ and the uniform SLLN, i.e.,

$$\sup_{(F, A) \in \Theta_k^*(5M)} |\Phi(F, A, P_n) - \Phi(F, A, P)| \xrightarrow{\text{a.s.}} 0,$$

the theorem is completely proven.

Let

$$\tilde{\theta}_n = \begin{cases} \hat{\theta}_n & \text{if } \hat{\theta}_n \in \Theta_k^*(5M) \\ \theta_* & \text{if } \hat{\theta}_n \notin \Theta_k^*(5M) \end{cases},$$

where $\theta_* \in \Theta_k^*(5M)$ is chosen to ensure $d(\theta_*, \Theta') > 0$. Then, for a sufficiently large n , $\tilde{\theta}_n = \hat{\theta}_n$ by Lemma 2, and the following condition is obtained

$$\limsup_n \left[\Phi(\tilde{\theta}_n, P_n) - \inf_{\theta \in \Theta'} \Phi(\theta, P_n) \right] \leq 0 \quad \text{a.s.}$$

Since $\limsup_n \Phi(\theta_0, P_n) = \Phi(\theta_0, P) (= m_k(P))$ for any fixed $\theta_0 \in \Theta'$,

$$\limsup_n \inf_{\theta \in \Theta'} \Phi(\theta, P_n) \leq \limsup_n \Phi(\theta_0, P_n) = m_k(P) \quad \text{a.s.}$$

Thus,

$$\begin{aligned} 0 &\geq \limsup_n \Phi(\tilde{\theta}_n, P_n) - \limsup_n \inf_{\theta \in \Theta'} \Phi(\theta, P_n) \\ &\geq \limsup_n \Phi(\tilde{\theta}_n, P_n) - m_k(P) \quad \text{a.s.} \end{aligned} \quad (11)$$

Let $\Theta_\epsilon^*(5M) := \{\theta \in \Theta_k^*(5M) \mid d(\theta, \Theta') \geq \epsilon\}$ for each $\epsilon > 0$. From the uniform SLLN,

$$\liminf_n \inf_{\theta \in \Theta_\epsilon^*(5M)} \Phi(\theta, P_n) \geq \inf_{\theta \in \Theta_\epsilon^*(5M)} \Phi(\theta, P) \quad \text{a.s.} \quad (12)$$

for all $\epsilon > 0$. An arbitrary $\epsilon > 0$ is selected. From Corollary 1 and the inequalities (11) and (12), we have

$$\liminf_n \inf_{\theta \in \Theta_\epsilon^*(5M)} \Phi(\theta, P_n) > \limsup_n \Phi(\tilde{\theta}_n, P_n) \quad \text{a.s.} \quad (13)$$

That is, for any $\omega \in \Omega$ satisfying the inequality (13), there exists $n_0 \in \mathbb{N}$ such that

$$\inf_{\theta \in \Theta_\epsilon^*(5M)} \Phi(\theta, P_n) > \Phi(\hat{\theta}_n, P_n) = \Phi(\tilde{\theta}_n, P_n)$$

for all $n \geq n_0$. Conversely, suppose that there exists $n \geq n_0$ such that $d(\hat{\theta}_n, \Theta') \geq \epsilon$. Then, we obtain

$$\inf_{\theta \in \Theta_\epsilon^*(5M)} \Phi(\theta, P_n) = \Phi(\hat{\theta}_n, P_n),$$

which is a contradiction. Thus, we obtain that $d(\hat{\theta}_n, \Theta') < \epsilon$ for all $n \geq n_0$. That is,

$$\lim_{n \rightarrow \infty} d(\hat{\theta}_n, \Theta') = 0 \quad \text{a.s.}$$

is proven. From the continuity of $\Phi(\cdot, P)$, the following is obtained:

$$\lim_{n \rightarrow \infty} m_k(P_n) = m_k(P) \quad \text{a.s.}$$

7. Selection of the number of dimensions

In RKM clustering, the numbers of clusters and dimensions, k and q , have to be appropriately determined such that the cluster result can be optimized. For determining the number of cluster, Wang (2010) proposed a new selection criterion based on clustering stability. This criterion can be applied for determining other turning parameters with some clustering method (e.g., Sun et al., 2012).

In this section, we propose a new simple criterion for determining the number of dimensions under given cluster number, which is not based on clustering stability. We also propose a consistent estimator of the criterion. Moreover, we illustrate the effectiveness of the criterion through numerical experiments.

7.1. New criterion for determining the number of dimensions

First, we define a variance ratio criterion for a population distribution P by

$$VR(q | P) := \inf_{(F, A) \in \Theta'} \frac{\int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x})}{\int \|A^T \mathbf{x} - A^T \boldsymbol{\mu}\|^2 P(d\mathbf{x})},$$

where $\boldsymbol{\mu} = \int \mathbf{x} P(d\mathbf{x})$.

Here, we assume that the population global optimal coefficient matrices are determined uniquely without the rotational indeterminacy of A , that is, there exists $(F_0, A_0) \in \Theta'$ such that for all $(F, A) \in \Theta'$ there exists $R \in \mathcal{O}(q)$ such that $A_0 = AR$. Let $(F, A), (F_*, A_*) \in \Theta'$ with $F \neq F_*$ or $A \neq A_*$. We have $\Phi(F, A, P) = \Phi(F_*, A_*, P)$ and $\int \|A^T \mathbf{x}\|^2 P(d\mathbf{x}) = \int \|A_*^T \mathbf{x}\|^2 P(d\mathbf{x})$. Since

$$\Phi(F, A, P) = \int \|\mathbf{x}\|^2 P(d\mathbf{x}) - \int \|A^T \mathbf{x}\|^2 P(d\mathbf{x}) + \int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x}),$$

we obtain

$$\frac{\int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x})}{\int \|A^T \mathbf{x} - A^T \boldsymbol{\mu}\|^2 P(d\mathbf{x})} = \frac{\int \min_{\mathbf{f}_* \in F_*} \|A_*^T \mathbf{x} - \mathbf{f}_*\|^2 P(d\mathbf{x})}{\int \|A_*^T \mathbf{x} - A_*^T \boldsymbol{\mu}\|^2 P(d\mathbf{x})}.$$

Unfortunately, we cannot obtain the value of this criterion since the population distribution is unknown. However, we can construct a consistent estimator of $VR(q | P)$. We define a estimator of $VR(q | P)$ by

$$\widehat{VR}(q | P_n) := \frac{\int \min_{\hat{\mathbf{f}}_n \in \hat{F}_n} \|\hat{A}_n^T \mathbf{x} - \hat{\mathbf{f}}_n\|^2 P_n(d\mathbf{x})}{\int \|\hat{A}_n^T \mathbf{x} - \hat{A}_n^T \boldsymbol{\mu}\|^2 P_n(d\mathbf{x})},$$

where $\hat{\theta}_n = (\hat{F}_n, \hat{A}_n)$. The following theorem gives the sufficient conditions of the strong consistency of the estimator $\widehat{VR}(q | P_n)$.

Theorem 2. *Suppose that $\int \phi(\|\mathbf{x}\|) P(d\mathbf{x}) < \infty$ and $m_1(P) > m_2(P) > \dots > m_k(P)$. Then,*

$$\int \|A^T \mathbf{x} - A^T \boldsymbol{\mu}\|^2 P(d\mathbf{x}) > 0 \quad \text{for all } (F, A) \in \Theta'$$

and

$$\lim_{n \rightarrow \infty} \widehat{VR}(q | P_n) = VR(q | P) \quad \text{a.s.}$$

Proof. Without loss of generality, we assume $\boldsymbol{\mu} = \mathbf{0}$. First, we prove

$$\int \|A^T(\mathbf{x} - \boldsymbol{\mu})\|^2 P(d\mathbf{x}) > 0 \quad \text{for all } (F, A) \in \Theta'.$$

Conversely, suppose that there exists $(F, A) \in \Theta'(k)$ such that $\int \|A^T \mathbf{x}\|^2 P(d\mathbf{x}) = 0$. Then, $\|A^T \mathbf{x}\|^2 = 0$ for all \mathbf{x} in the support of P . Since

$$\Phi(F, A, P) = \int \|\mathbf{x} - AA^T \mathbf{x}\|^2 P(d\mathbf{x}) + \int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x}),$$

F must contain zero. Let $F_0 := \{\mathbf{0}\} \in \mathcal{R}_1$ and then $m_k(P) = \Phi(F_0, A, P) \geq m_1(P)$. This is a contradiction.

Next, we prove the consistency of $\widehat{VR}(q | P_n)$. From Theorem 1, we have

$$\lim_{n \rightarrow \infty} d(\hat{\theta}_n, \Theta') = 0 \quad \text{a.s.}$$

In the similar way as the proof of the uniform SLLN (6), we obtain

$$\lim_{n \rightarrow \infty} \sup_{A \in \mathcal{O}(p \times q)} \left| \int \|A^T \mathbf{x}\|^2 P_n(d\mathbf{x}) - \int \|A^T \mathbf{x}\|^2 P(d\mathbf{x}) \right| = 0 \quad \text{a.s.} \quad (14)$$

and

$$\lim_{n \rightarrow \infty} \sup_{(F, A) \in \Theta} \left| \int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P_n(d\mathbf{x}) - \int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x}) \right| = 0 \quad \text{a.s.} \quad (15)$$

Let $\hat{\theta}_n = (\hat{F}_n, \hat{A}_n)$ and $(F, A) \in \Theta'$. We have

$$\lim_{n \rightarrow \infty} \int \|\hat{A}_n^T \mathbf{x}\|^2 P_n(d\mathbf{x}) = \int \|A^T \mathbf{x}\|^2 P(d\mathbf{x}) \quad \text{a.s.}$$

and

$$\lim_{n \rightarrow \infty} \int \min_{\hat{\mathbf{f}}_n \in \hat{F}_n} \|\hat{A}_n^T \mathbf{x} - \hat{\mathbf{f}}_n\|^2 P_n(d\mathbf{x}) = \int \min_{\mathbf{f} \in F} \|A^T \mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x}) \quad \text{a.s.}$$

Therefore, we obtain

$$\lim_{n \rightarrow \infty} \widehat{VR}(q \mid P_n) = VR(q \mid P) \quad \text{a.s.}$$

□

If the number of dimensions is determined larger than the optimal one, the subspace of RKM may be influenced from noise variables which do not have cluster structure. Let q_* be the optimal number of dimensions. Define $VR(0 \mid P) := 0$ and $VR(q \mid P) := VR(q - 1 \mid P)$ for $q = \min\{k - 1, p\}$. Forward difference at q_* , $\Delta_+(q) := VR(q_* + 1 \mid P) - VR(q_* \mid P)$, may be quite larger than backward difference at q_* , $\Delta_-(q) := VR(q_* \mid P) - VR(q_* - 1 \mid P)$. That is, for the optimal number of dimensions q_* , second order central difference at q_* , $\Delta_2(q_*) := VR(q_* + 1 \mid P) - 2VR(q_* \mid P) + VR(q_* - 1 \mid P)$, may be larger than second order central difference at q ($q \neq q_*$). For example, we may estimate the optimal number of dimensions by

$$\hat{q} := \arg \max_q \widehat{\Delta}_2(q),$$

where $\widehat{\Delta}_2(q) := \widehat{VR}(q + 1 \mid P) - 2\widehat{VR}(q \mid P) + \widehat{VR}(q - 1 \mid P)$.

7.2. Numerical experiments

In this subsection, we examine the effectiveness of the criterion through numerical experiments. Let K be the number of clusters, q be the number of dimensions of the low dimensional space, p_1 be the number of the informative variables, p_2 be the number of the correlated noise variables, and p_3 be the number of the independent noise variables. Denote $O_{p \times q}$ be the $p \times q$ zero matrix. The $p_1 \times q$ column wise orthogonal matrix is generated randomly, say A_* . K cluster centers in low-dimensional space are independently generated from the q -dimensional uniform distribution on $[-15, 15]^q$, say \mathbf{f}_k ($k = 1, \dots, K$). Cluster indicators are independently generated from the multinomial distribution for K trials with equal probabilities, say $\mathbf{u}_i = (u_{i1}, \dots, u_{iK})$ ($i = 1, \dots, n$). Set $A = [A_*^T, O_{q \times (p_2 + p_3)}]^T$, $\Sigma_p = (\sigma_{ij})_{p_2 \times p_2}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.25$ ($i \neq j$), and

$$\Sigma_p = \begin{bmatrix} I_{p_1} & O_{p_1 \times p_2} & O_{p_1 \times p_3} \\ O_{p_2 \times p_1} & \Sigma_{p_2} & O_{p_2 \times p_3} \\ O_{p_3 \times p_1} & O_{p_3 \times p_2} & I_{p_3} \end{bmatrix}.$$

The simulated data of n observations, $\mathbf{x}_i \in \mathbb{R}^p$ ($i = 1, \dots, n$), are generated as

$$\mathbf{x}_i = \sum_{k=1}^K u_{ik} (A \mathbf{f}_k + \boldsymbol{\epsilon}_{ik}),$$

where $\boldsymbol{\epsilon}_{ik}$ are generated from the p -dimensional normal distribution $N(\mathbf{0}, \Sigma_p)$. Let $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and Z be the normalized data matrix with zero means and unit variances.

Here, we set $K = 8$, $n = 400$, $q = 2$ or 3 and $p_1 = p_2 = p_3 = 5$ or 10 . We make 1000 data sets for each setting, respectively. Figure 3 shows hidden cluster structure XA of the one of data set with setting $n = 400$, $q = 2$, and $p_1 = p_2 = p_3 = 5$. Figure 4 shows the first two principal components of PCA for Z , which is the same data set of Figure 3 and also shows that the first two principal components do not reflect the cluster structure. Moreover, Figure 5 shows the subspace of RKM with $q = 2$ for Z , which is the same data set of Figure 3. Figure 6 shows the adjusted rand indexes (ARI), which is proposed by Hubert and Arabie (1985), of RKM clustering with each number of dimensions of subspace. In Figure 6, we can see that the number of dimensions of the subspace is quite important to the clustering result. Figure 7 and 8

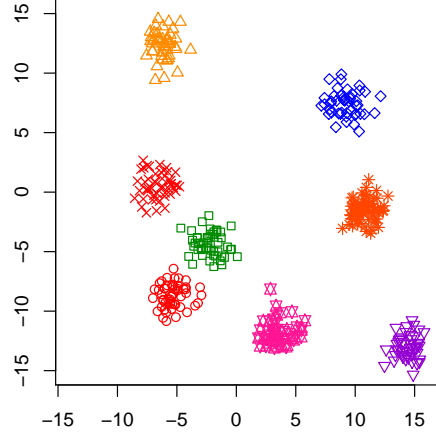


FIG 3. Hidden cluster structure XA of the one of data set with setting $n = 400$, $q = 2$, and $p_1 = p_2 = p_3 = 5$.

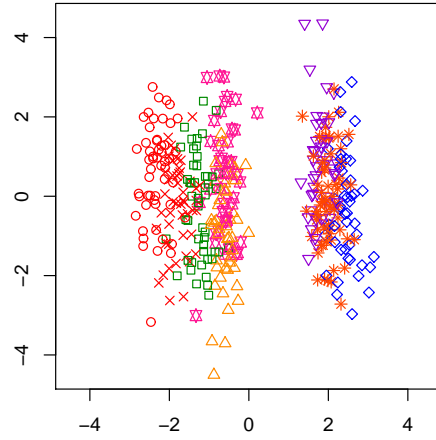


FIG 4. First two dimensions of the principal scores of PCA for Z , which is the same data set of Figure 3 (ARI of the tandem clustering with first two principal scores is 0.26).

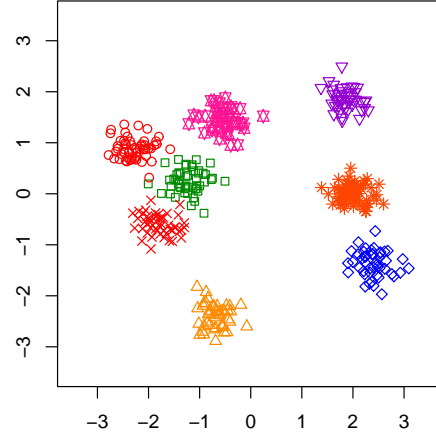


FIG 5. The subspace of RKM for Z , which is the same data set of Figure 3 (ARI of the RKM clustering with $q = 2$ is 0.99).

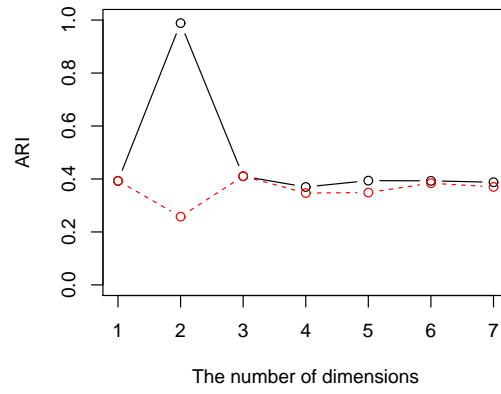


FIG 6. ARI scores of RKM and tandem clustering with $q = 1, 2, \dots, 7$ for Z , which is the same data set of Figure 3. Solid line is corresponded to ARI scores of RKM clustering and dash line is corresponded to ARI scores of tandem clustering.

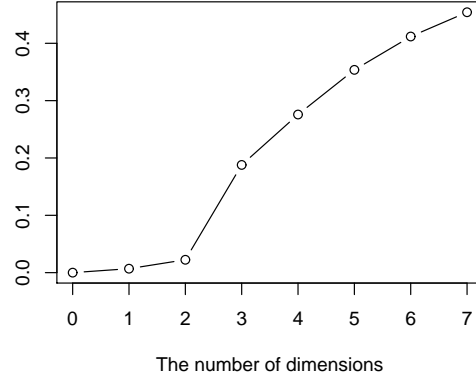


FIG 7. $\widehat{VR}(q)$ scores of RKM with $q = 1, 2, \dots, 7$ for Z , which is the same data set of Figure 3.

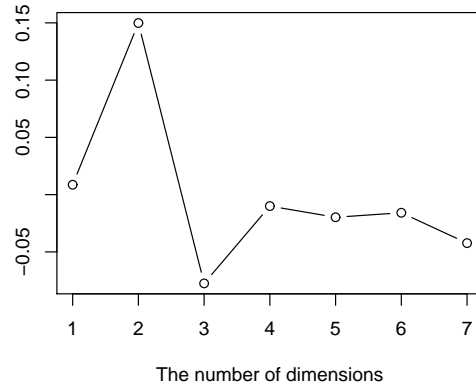


FIG 8. $\widehat{\Delta}_2(q)$ scores of RKM with $q = 1, 2, \dots, 7$ for Z , which is the same data set of Figure 3.

TABLE 1
Agreement rates with each setting for 1000 data sets.

q	$p_1 = p_2 = p_3$	agreement rate
2	5	0.84 (837/1000)
2	10	0.95 (947/1000)
3	5	0.73 (726/1000)
3	10	0.89 (890/1000)

show that $\widehat{VR}(q)$ and $\widehat{\Delta}_2(q)$ are useful for estimating the optimal number of dimensions.

Indeed, Table 1 shows the agreement rates, of the choices by \hat{q} and the optimal number $q_* := \arg \max_q ARI(q)$, with each setting for 1000 data sets.

8. Conclusion

This paper proves the strong consistency of RKM clusterings under i.i.d. sampling on the basis of the proof for the conventional k -means clustering provided by Pollard (1981). Since our proof is based on the usual Blum-DeHardt uniform SLLN which requires only stationarity and ergodicity (e.g., Peskir, 2000), we can obtain the same results for a stationary ergodic process.

Under the i.i.d. condition, we can derive the rate of convergence for the convergence of the empirically optimal clustering scheme if the support of the population distribution is bounded; that is, $P(\|\mathbf{X}_1\|^2 \leq B) = 1$ for some $B > 0$. From Theorem 1 in Linder et al. (1994), for all $\epsilon > 0$ and $n(\epsilon/8B)^2 \geq 2$ we can obviously obtain

$$\begin{aligned}
P[|m_k(P_n) - m_k(P)| \geq \epsilon] &\leq 2P\left[\sup_{\theta \in \Xi_k} |\Phi(\theta, P_n) - \Phi(\theta, P)| \geq \epsilon\right] \\
&\leq 2P\left[\sup_{F \in \mathcal{R}_k^{(p)}} |KM(F, P_n) - KM(F, P)| \geq \epsilon\right] \\
&\leq 8(2n)^{k(p+1)} \exp\left(-\frac{n\epsilon^2}{512B^2}\right),
\end{aligned}$$

where $\phi(r) = r^2$, $\mathcal{R}_k^{(p)} := \{R \subset \mathbb{R}^p \mid \#(R) \leq k\}$, and $KM(F, P) := \int \min_{\mathbf{f} \in F} \|\mathbf{x} - \mathbf{f}\|^2 P(d\mathbf{x})$.

Considering the relationship between the conventional k -means clustering and RKM clustering, the results presented in this paper are applicable to

the conventional k -means clustering. The related methods of RKM clustering include factorial k -means (FKM) clustering proposed by Vichi & Kiers (2001). In Terada (2013), the strong consistency of FKM clusterings under i.i.d. sampling (or for a stationary ergodic process) has been proven. The form of sufficient conditions for the strong consistency of FKM clustering is similar to the case of RKM clusterings. Moreover, the new simple criterion for determining the number of dimensions under given cluster number and the consistent estimator of the criterion have been proposed. Through numerical experiments, the effectiveness of the criterion has been illustrated.

Future studies in this regard will examine the rate of convergence of estimators of RKM clustering and will propose the criterion required to determine the number of clusters.

References

- [1] ABRAHAM, C., CORNILLON, P.A., MATZNER-LØBER, E. & MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Statist.* **30**, 581–595. [MR2002229](#)
- [2] ARABIE, P. & HUBERT, L. (1994). Cluster Analysis in Marketing Research. In: Bagozzi, R.P. (Eds.), *Advanced Methods of Marketing Research*. Oxford, Blackwell, 160–189.
- [3] CHANG, W. (1994). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*. **32**, 267–275. [MR0770316](#)
- [4] DE SOETE, G. & CARROLL, J. D. (1994). K -means clustering in a low-dimensional Euclidean space. In: Diday, E., et al. (Eds.), *New Approaches in Data Analysis*. Springer, Heidelberg, 212–219.
- [5] DEHARDT, J. (1971). Generalizations of the Glivenko-Cantelli Theorem. *Ann. Math. Statist.* **42**, 2050–2055. [MR297000](#)
- [6] GARCÍA-ESCUADERO, L.A., GORDALIZA, A. & MATRÁN, C (1999). A central limit theorem for multivariate generalized trimmed k -means. *Ann. Statist.* **27**, 1061–1079. [MR1724041](#)
- [7] HUBERT, L. & ARABIE, P. (1985). Comparing partitions. *J. Classification*. **2**, 193–218.
- [8] LINDER, T., LUGOSI, G. & ZEGGER, K. (1994). Rates of convergence in the source coding theorem, empirical quantizer design, and universal lossy source coding. *IEEE Trans. Inform. Theory*. **40**, 1728–1740. [MR1322387](#)

- [9] PESKIR, G. (2000). *From Uniform Laws of Large Numbers to Uniform Ergodic Theorems*. Univ. Aarhus, Dept. Mathematics. [MR1805157](#)
- [10] PFANZAGL, J. (1994). *Parametric Statistical Theory*. de Gruyter, Berlin. [MR1291393](#)
- [11] POLLARD, D. (1981). Strong consistency of k -means clustering. *Ann. Statist.* **9**, 135–140. [MR0600539](#)
- [12] POLLARD, D. (1982). A central limit theorem for k -means clustering. *Ann. Probab.* **10**, 919–926. [MR0672292](#)
- [13] TERADA, Y. (2013). Strong consistency of factorial k -means clustering. *arXiv*.
- [14] TIMMERMAN, M.E., CEULEMANS, E., KIERS, H.A.L. & VICHI, M. (2010). Factorial and reduced K -means reconsidered. *Comput. Statist. Data Anal.* **54**, 1858–1871. [MR2608979](#)
- [15] VICHI, M. & KIERS, H.A.L. (2001). Factorial k -means analysis for two-way data. *Comput. Statist. Data Anal.* **37**, 49–64. [MR1862479](#)
- [16] VAN DER VAART, A. (1998). *Asymptotic Statistics*. Cambridge Univ. Press. [MR1652247](#)
- [17] VON LUXBURG, U., BELKIN, M. & BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36**, 555–586. [MR2396807](#)
- [18] WANG, J. (2010). Consistent Selection of the Number of Clusters via Cross Validation. *Biometrika.* **97**, 893–904. [MR2396807](#)

Appendix A: The existence of Θ'

The existence of the minimum points of $\Phi(\cdot, P)$ are proven.

Lemma 3. *Suppose that $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \infty$. There exists $M > 0$ such that, for all $F' \in \mathcal{R}_k$ satisfying $F' \cap B_q(M) = \emptyset$,*

$$\inf_{A \in \mathcal{O}(p \times q)} \Phi(F', A, P) > \inf_{\theta \in \Theta_k^*(M)} \Phi(\theta, P).$$

Proof. Argue by contradiction, suppose that for any $M > 0$ there exists $F' \in \mathcal{R}_k$ such that $F' \cap B_q(M) = \emptyset$ and

$$\inf_{A \in \mathcal{O}(p \times q)} \Phi(F', A, P) \leq \inf_{\theta \in \Theta_k^*(M)} \Phi(\theta, P). \quad (16)$$

Select an $r > 0$ such that the ball $B_p(r)$ has a positive P -measure, i.e., $P(B_p(r)) > 0$. A sufficient large M is selected such that $M > r$ and inequality

(8) is satisfied. From the inequality (16),

$$\begin{aligned} \int \phi(\|\mathbf{x}\|)P(\mathbf{x}) &\geq \inf_{\theta \in \Theta_k^*(M)} \Phi(\theta, P) \geq \inf_{A \in \mathcal{O}(p \times q)} \Phi(F', A, P) \\ &\geq \phi(M - r)P(B_p(r)). \end{aligned}$$

This is a contradiction. \square

Lemma 4. *Suppose that $\int \phi(\|\mathbf{x}\|)P(d\mathbf{x}) < \infty$ and $m_j(P) > m_k(P)$ for $j = 1, 2, \dots, k-1$. There exists $M > 0$ such that, for any $F' \in \mathcal{R}_k$ satisfying $F' \not\subset B_q(5M)$,*

$$\inf_{A \in \mathcal{O}(p \times q)} \Phi(F', A, P) > \inf_{\theta \in \Theta_k^*(5M)} \Phi(\theta, P).$$

Proof. Select a sufficient large value $M > 0$ to satisfy the inequalities (8) and (9). To obtain a contradiction, suppose that for all $M > 0$ there exists $F' \in \mathcal{R}_k$ satisfying $F' \not\subset B_q(5M)$ and

$$\inf_{A \in \mathcal{O}(p \times q)} \Phi(F', A, P) \leq \inf_{\theta \in \Theta_k^*(5M)} \Phi(\theta, P).$$

Let \mathcal{R}'_k be the set of such F' so that

$$m_k(P) = \inf_{\theta \in \mathcal{R}'_k \times \mathcal{O}(p \times q)} \Phi(\theta, P).$$

From Lemma 3, each $F' \in \mathcal{R}'_k$ includes at least one element in $B_q(M)$, say \mathbf{f}_1 .

For any \mathbf{x} satisfying $\|\mathbf{x}\| < 2M$ and any $A \in \mathcal{O}(p \times q)$,

$$\|\mathbf{x} - A\mathbf{f}\| > 3M \quad \text{for all } \mathbf{f} \notin B_q(5M)$$

and

$$\|\mathbf{x} - A\mathbf{g}\| < 3M \quad \text{for all } \mathbf{g} \in B_q(M).$$

Let F^* denote the set obtained by deleting all elements outside $B_q(5M)$ from F' . Then,

$$\int_{\|\mathbf{x}\| < 2M} \min_{\mathbf{f} \in F'} \phi(\|\mathbf{x} - A\mathbf{f}\|)P(d\mathbf{x}) = \int_{\|\mathbf{x}\| < 2M} \min_{\mathbf{f} \in F^*} \phi(\|\mathbf{x} - A\mathbf{f}\|)P(d\mathbf{x}).$$

Since $\int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x} - A\mathbf{f}_1\|)P(d\mathbf{x}) \leq \lambda \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\|)P(d\mathbf{x})$, we obtain

$$\begin{aligned} & \Phi(F', A, P) + \lambda \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x}\|)P(d\mathbf{x}) \\ & \geq \int_{\|\mathbf{x}\| < 2M} \min_{\mathbf{f} \in F^*} \phi(\|\mathbf{x} - A\mathbf{f}\|)P(d\mathbf{x}) + \int_{\|\mathbf{x}\| \geq 2M} \phi(\|\mathbf{x} - A\mathbf{f}_1\|)P(d\mathbf{x}) \\ & \geq \Phi(F^*, A, P) \geq m_{k-1}(P) \end{aligned}$$

for all $A \in \mathcal{O}(p \times q)$. Therefore, we obtain

$$m_k(P) + \epsilon \geq m_{k-1}(P).$$

This contradicts $m_k(P) + \epsilon < m_{k-1}(P)$. \square

We will denote the essential parameter space by Θ_k ; that is, $\Theta_k := \Theta_k^*(5M)$. By Lemma 4,

$$\inf_{\theta \in \Xi_k} \Phi(\theta, P) = \inf_{\theta \in \Theta_k} \Phi(\theta, P)$$

and there is no $\theta \in (\mathcal{R}_k \setminus \mathcal{R}_k^*(5M)) \times \mathcal{O}(p \times q)$ satisfying $m_k(P) = \Phi(\theta, P)$.

Proof of Proposition 3. First, it is proven that there exists a sequence $\{\theta_n\}_{n \in \mathbb{N}}$ in Θ_k such that $\Phi(\theta_n, P) \rightarrow m_k(P)$ as $n \rightarrow \infty$. Let $C := \{\Phi(\theta, P) \mid \theta \in \Theta_k\}$ and $m_k(P) = \inf C$. For all $x > m_k(P)$, there exists $c < x$ in C . Write $x_n := m_k(P) + 1/n$ and $C_n := \{c \in C \mid c < x_n\}$. Let $\mathfrak{P}(C)$ be the power set of C . From the axiom of choice, there exists a function $f : \mathfrak{P}(C) \setminus \{\emptyset\} \rightarrow C$ such that $f(B) \in B$ for all $B \in \mathfrak{P}(C) \setminus \{\emptyset\}$. Let $c_n := f(C_n)$ and $x_n > c_n \geq m_k(P)$. Thus, $c_n \rightarrow m_k(P)$ as $n \rightarrow \infty$. Using the axiom of choice, a sequence $\{\theta_n\}_{n \in \mathbb{N}}$ can be selected such that $\Phi(\theta_n, P) \rightarrow m_k(P)$ as $n \rightarrow \infty$.

From the compactness of Θ_k , there exists a convergent subsequence of $\{\theta_n\}_{n \in \mathbb{N}}$, say $\{\theta_{m_i}\}_{i \in \mathbb{N}}$. Let $\theta^* \in \Theta_k$ denote the limit of such subsequence, that is, $\theta_{m_i} \rightarrow \theta^*$ as $i \rightarrow \infty$. Because $\Phi(\cdot, P)$ is continuous on Θ_k , $\Phi(\theta^*, P) = m_k(P)$. That is, $\Theta' \neq \emptyset$. \square

The next corollary ensures the identification condition for $\Phi(\cdot, P)$.

Corollary 1. *Let $\Theta' := \{\theta_k \in \Theta_k \mid \Phi(\theta_k, P) = m_k(P)\}$. Assume the assumptions of Lemma 4. Then,*

$$\inf_{\theta \in \Theta_k : d(\theta, \Theta') \geq \epsilon} \Phi(\theta, P) > \inf_{\theta \in \Theta'} \Phi(\theta, P) \quad \text{for all } \epsilon > 0.$$

Proof. Let $\Theta_\epsilon := \{\theta \in \Theta_k \mid d(\theta, \Theta') \geq \epsilon\}$. To obtain a contradiction, suppose that there exists $\epsilon > 0$ such that $\inf_{\theta \in \Theta_\epsilon} \Phi(\theta, P) = \inf_{\theta \in \Theta'} \Phi(\theta, P)$. Like in the proof of Proposition 3, there exists a sequence $\{\theta_n\}_{n \in \mathbb{N}}$ on Θ_ϵ satisfying $\Phi(\theta_n, P) \rightarrow m_k(P)$ as $n \rightarrow \infty$. From the compactness of Θ_k , there exists a convergent subsequence of $\{\theta_n\}_{n \in \mathbb{N}}$, say $\{\theta_{m_i}\}_{i \in \mathbb{N}}$. Let $\theta^* \in \Theta_k$ denote the limit of such subsequence and $\Phi(\theta^*, P) = m_k(P)$, that is, $\theta^* \in \Theta'$. On the other hand, $d(\theta_{m_i}, \theta^*) < \epsilon$ for sufficiently large $i \in \mathbb{N}$ because $\theta_{m_i} \rightarrow \theta^*$ as $i \rightarrow \infty$. Thus, $\theta_{m_i} \notin \Theta_\epsilon$ for sufficiently large $i \in \mathbb{N}$. This is a contradiction. \square